Linköping Electronic Articles in
Computer and Information Science
Vol. 5(2000): nr 2

Simple Pre- and Post-Pruning Techniques for Large Conceptual Clustering Structures

Guy Mineau, Akshay Bissoon, and Robert Godin

Linköping University Electronic Press
Linköping, Sweden

http://www.ep.liu.se/ea/cis/2000/002/

Posted on March 8,
formally published on December 5, 2000
by Linköping University Electronic Press
581 83 Linköping, Sweden

Linköping Electronic Articles in
Computer and Information Science
ISSN 1401-9841
Series editor: Erik Sandewall

Typeset by the author using LaTeX
Formatted using étendu *style*

Recommended citation:
<Author>. <Title>. Linköping Electronic Articles in Computer and Information Science, Vol. 5(2000): nr 2. http://www.ep.liu.se/ea/cis/2000/002/. *December 5, 2000.*

This URL will also contain a link to the author's home page.

Abstract

In (Godin et al, 1995a) we proposed an incremental conceptual clustering algorithm, derived from lattice theory (Godin et al., 1995b), which is fast to compute (Mineau & Godin, 1995). This algorithm is especially useful when dealing with large data or knowledge bases, making classification structures available to large size applications like those found in industrial settings.

However, in order to be applicable on large data sets, the analysis component of the algorithm had to be simplified: the thoroguh comparison of objects normally needed to fully justify the formation of classes had to be cut down. Of course, from less analysis results classes which carry les semantics, or which should not have been formed in the first place. Consequently, some classess are useless in terms of the information needs of the applications that will later on interact with the data. Pruning techniques are thus needed to eliminate these classes and simplify the classification structure.

However, since these classification structures are huge, the pruning techniques themselves must be simle so that they can be applied in reasonable time on large classification structures. This paper presents three such techniques: one is based on the definition of constraints over the generalization language, the other two are based on discrimination metrics applied on links between classes or on the classes themselves. Because the first technique is applied before the classification structure is built, it is called the *pre-pruning* technique, while the other two are called *post-pruning* techniques.

Authors' addresses

GM, AB: Department of Computer Science
Université Laval
Quebec City, Québec
Canada, G1K 7P4
E-mail: `mineau@ift.ulaval.ca`

RG: Department of Computer Science
Université du Québec à Montréal
Montréal, Québec
Canada, H3C 3P8
E-mail: `godin.robert@uqam.ca`

For technical reasons, this article is stored as two separate computer files, labelled "Part I" and "Part II", respectively. Please replace the present page by a printout of the "Part II" file.